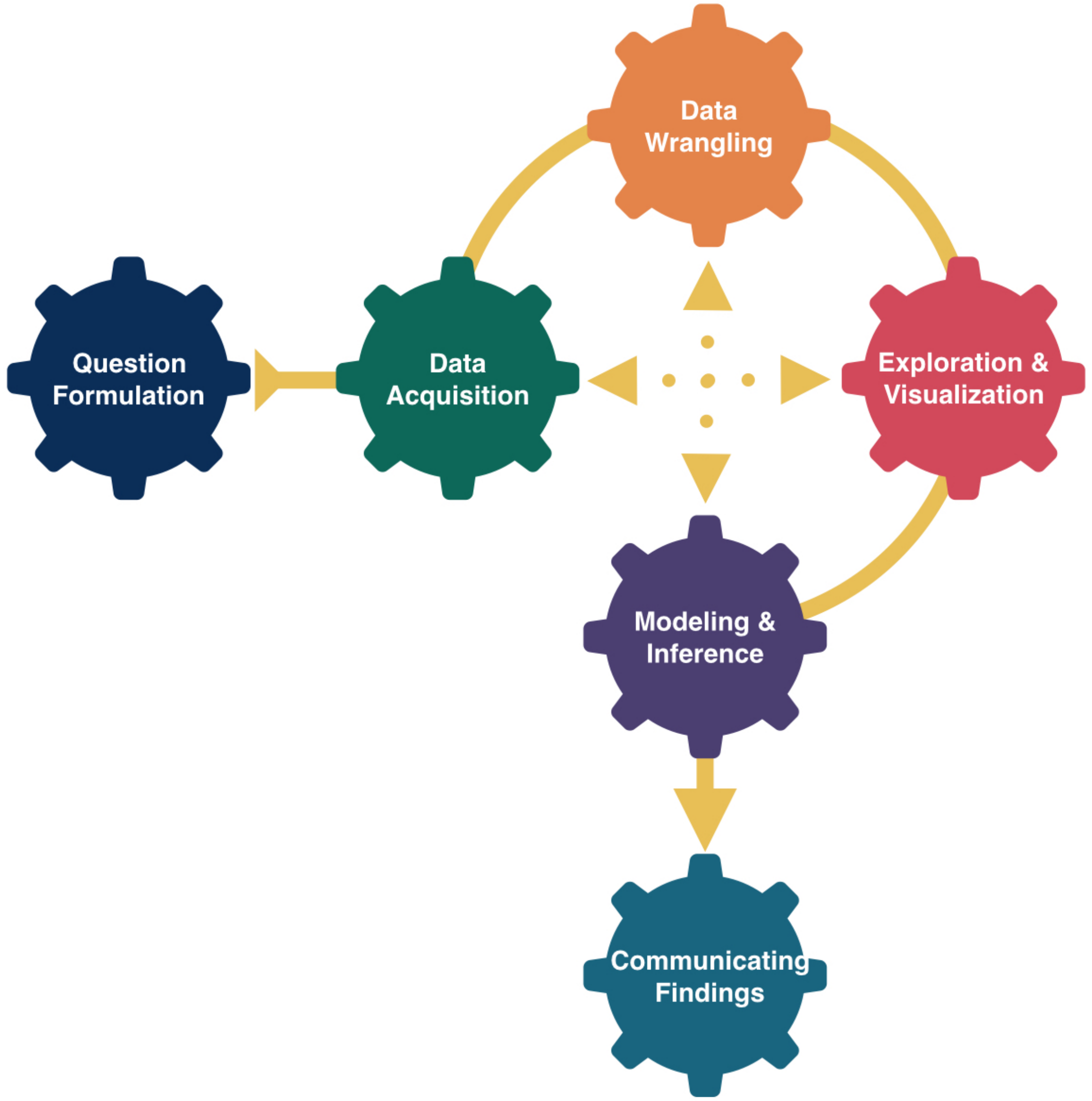


Hypothesis Testing

Kelly McConville

Stat 100

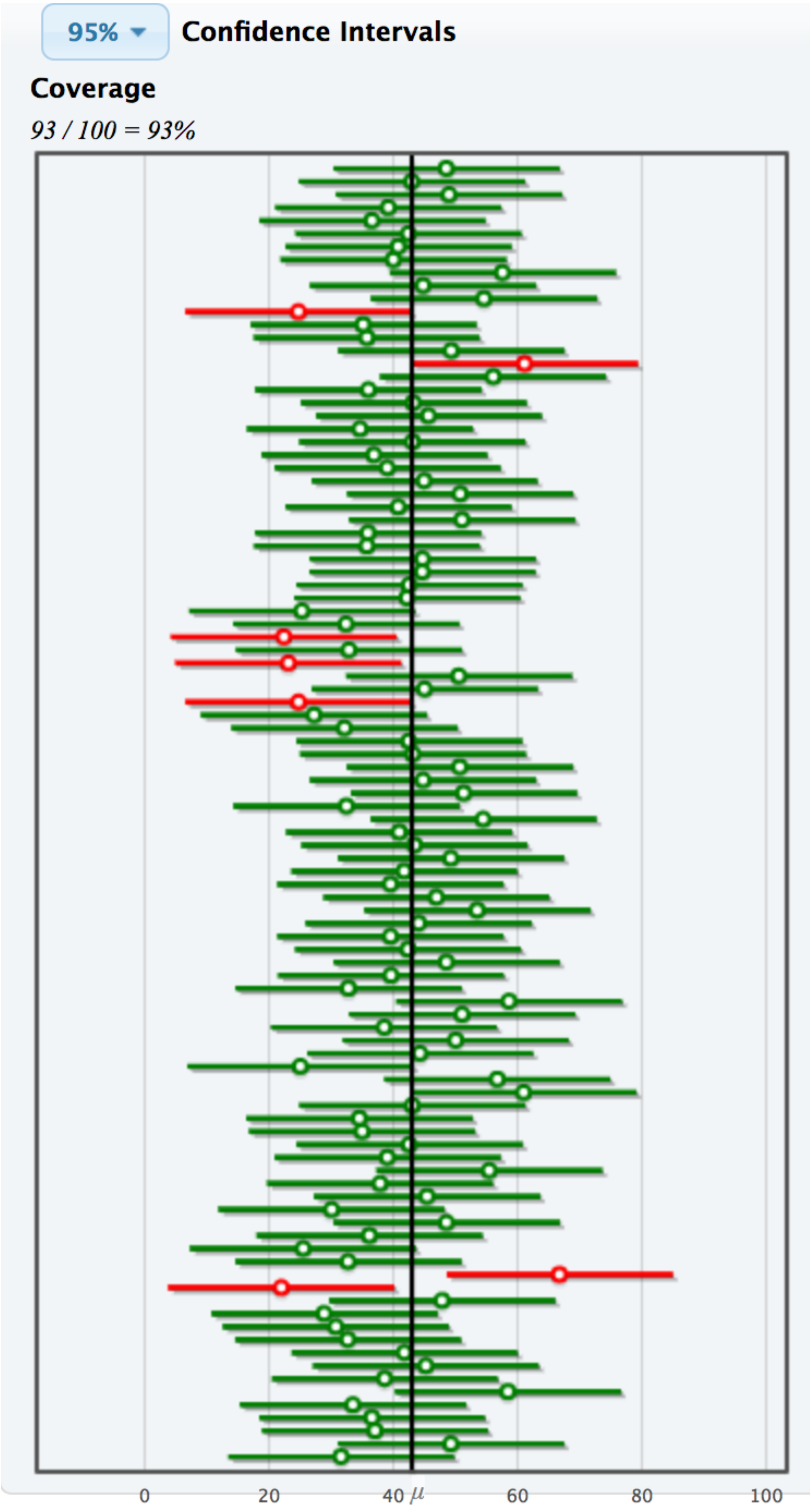
Week 9 | Fall 2023

Announcements

- Don’t forget that the midterm exam rewrites are due on Thursday at 5pm on Gradescope.
    - Make sure to use the Quarto doc in the Midterm Exam (Rewrites) project on Posit Cloud.
- 🎉 We are now accepting Course Assistant/Teaching Fellow applications for Stat 100 for next semester. To apply, fill out this application by **Nov 15th**.
    - About 10-12 hours of work per week.
    - Primary responsibilities: Attend weekly team meetings, lead a discussion section, hold office hours, grade assessments.

Goals for Today

- Confidence interval interpretations
- Set-up the structure of **hypothesis testing**
- Determine if Harvard students have ESP!



What do we mean by confidence?

- Confidence level = success rate of the method under **repeated sampling**
- How do I know if my ONE CI successfully contains the true value of the parameter?
- As we increase the **confidence level**, what happens to the width of the interval?
- As we increase the **sample size**, what happens to the width of the interval?
- As we increase the **number of bootstrap samples** we take, what happens to the width of the interval?

Interpreting Confidence Intervals

Example: Estimating average household income before taxes in the US

SE Method Formula:

$$\text{statistic} \pm \text{ME}$$

```
# A tibble: 1 × 3
      ME  lower  upper
   <dbl>  <dbl>  <dbl>
1 1929. 60551. 64409.
```

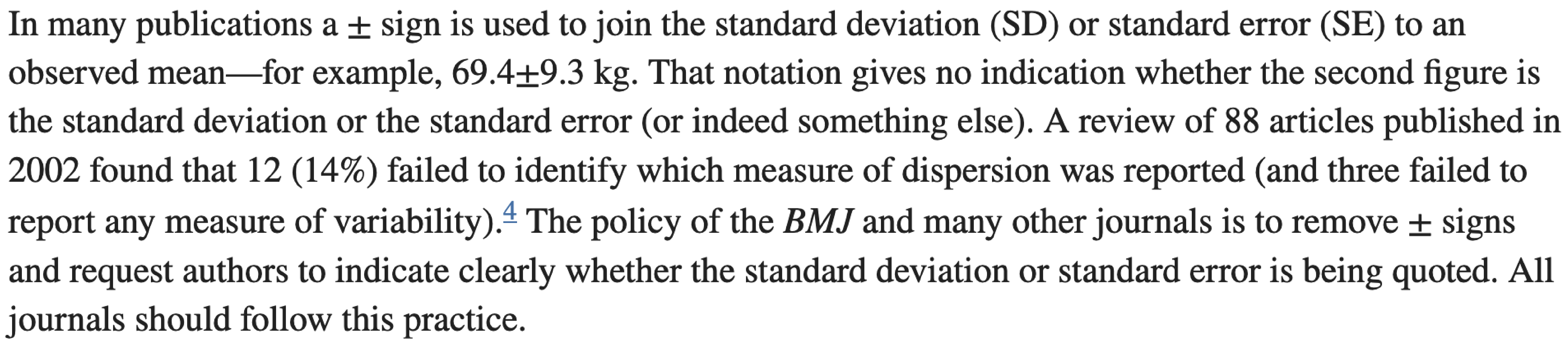
"The margin of [sampling] error can be described as the 'penalty' in precision for not talking to everyone in a given population. It describes the range that an answer likely falls between if the survey had reached everyone in a population, instead of just a sample of that population." – Courtney Kennedy, Director of Survey Research at Pew Research Center

CI = interval of **plausible** values for the **parameter**

Safe interpretation: I am P% confident that {insert what the parameter represents in context} is between {insert lower bound} and {insert upper bound}.

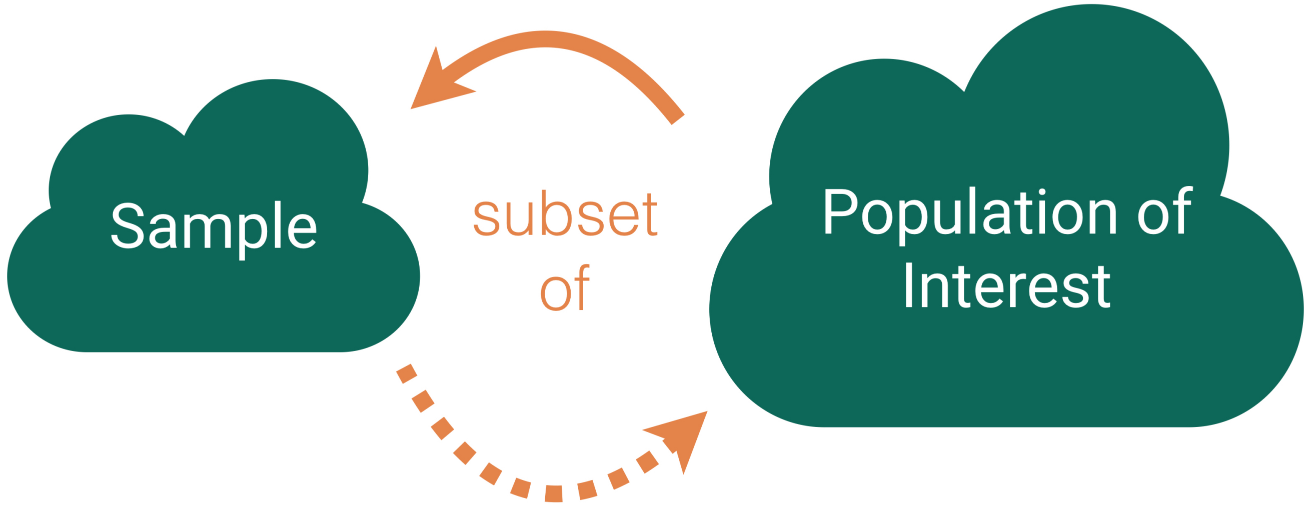
Caution: Confidence intervals in the wild

Statement in an article for The BMJ (British Medical Journal):



In many publications a ± sign is used to join the standard deviation (SD) or standard error (SE) to an observed mean—for example, 69.4±9.3 kg. That notation gives no indication whether the second figure is the standard deviation or the standard error (or indeed something else). A review of 88 articles published in 2002 found that 12 (14%) failed to identify which measure of dispersion was reported (and three failed to report any measure of variability).[4] The policy of the *BMJ* and many other journals is to remove ± signs and request authors to indicate clearly whether the standard deviation or standard error is being quoted. All journals should follow this practice.

Statistical Inference



Goal: Draw conclusions about the population based on a sample.

Main Flavors:

- Estimating numerical quantities.
- Testing conjectures.

Example: Does Extrasensory Perception (ESP) exist?

Daryl Bem and Ben Honorton

Bem and Honorton conducted extrasensory perception studies:

- A “sender” randomly chooses an object out of 4 possible objects and sends that information to a “receiver”.
- The “receiver” is then given a set of 4 possible objects and they must decide which one most resembles the object sent to them.

Out of 329 trials, the “receivers” correctly identified the object 106 times.

ESP Example

Let’s consider the following questions:

a. If ESP does not exist and the “receivers” are guessing, how often would we expect them to be correct?

b. For each sample (set of 329 trials), do we expect the proportion of correct guesses to be equal? Why or why not?

c. Is it possible to randomly guess correctly 106 out of 329 times (i.e., 32% of the time)?

d. How unusual is it to guess correctly 106 out of 329 times if ESP doesn’t exist?

To help us answer d., we need a **sampling distribution** for the sample proportion where **we assume the “receivers” were purely guessing!**

Sampling Distribution of a Statistic

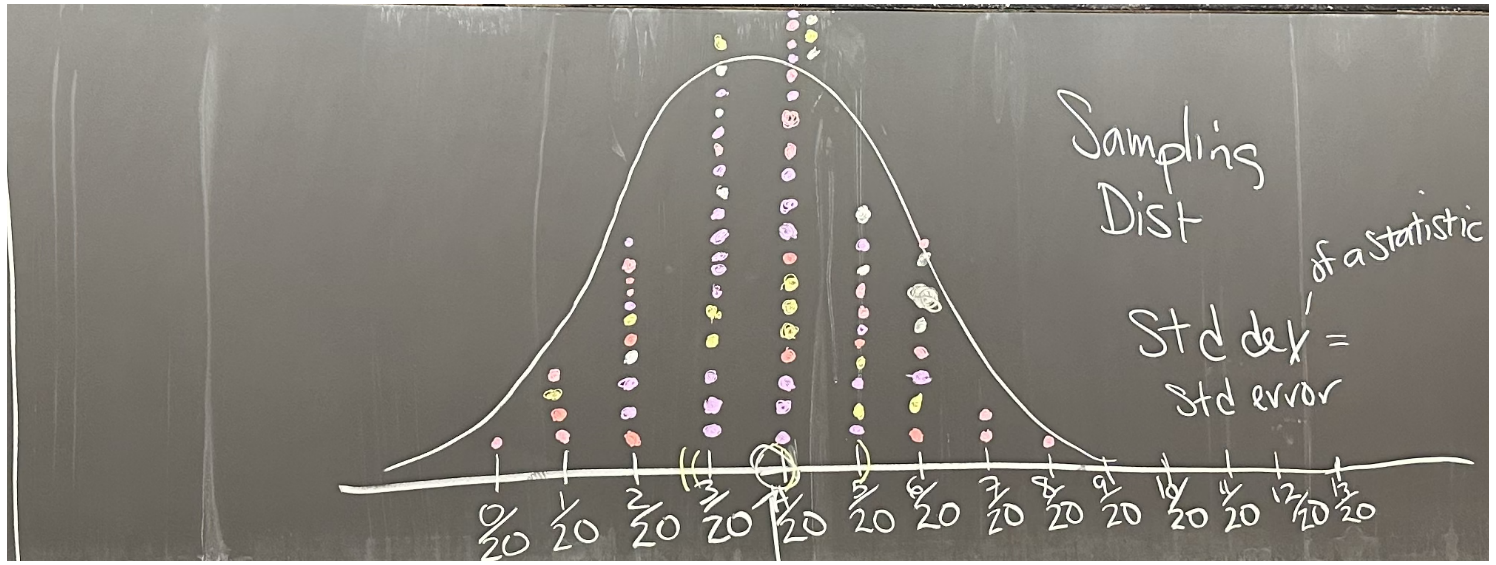
Steps for (Approximate) Distribution:

1. Decide on a sample size, n.
2. Randomly select a sample of size n from the population.
3. Compute the sample statistic.
4. Put the sample back in.
5. Repeat Steps (2) - (4) many (1000+) times.

Sampling Distribution of a Statistic

Steps for (Approximate) Distribution:

1. Decide on a sample size, n.
2. Randomly select a sample of size n from the population.
3. Compute the sample statistic.
4. Put the sample back in.
5. Repeat Steps (2) - (4) many (1000+) times.



Sampling Distribution Under No ESP

Steps for (Approximate) Distribution:

1. Decide on a sample size, n.

2. Randomly select a sample of size n from the population.

```r
library(mosaic)
rflip(n = 329, prob = 0.25)
```

```
Flipping 329 coins [ Prob(Heads) = 0.25 ] ...

H H H T T H H T T T T T T T H T T H H T T T T T T H T H H T H T T T T T
T T H H T T T T T T H T T T T T T T T H T T T T T T T T H T T T T T T T
T H H T T T H T T H T T T T T H T T H H H H T T T T T T T T T T T T T H
T T T T T H T T T T T T T T T T H T T H H T T T T T H T T T T H T T T T
T T T T H T T T T T T T H H T H T T H T T T H T T H T T T T T T T T T H
H T T T T T T T H H T T T T H T T T H T H H H H T T T T H H T T H T T T
H H T H T T T T H H T T H T H T H H H H T H T T T T H T T H H T T T T T
H T T T T T T H H T T T T T T H T T T H T T T T T T T T T T T T H T T T
T T H T T H T T T T H H H H T T T T T T T H H T T T H T T H T T T T H T
T H T H T

Number of Heads: 87 [Proportion Heads: 0.264437689969605]
```

Sampling Distribution Under No ESP

3. Compute the sample statistic.

```r
rflip(n = 329, prob = 0.25, summarize = TRUE)
```

```
    n heads tails prob
1 329    94   235 0.25
```

4. Put the sample back in.

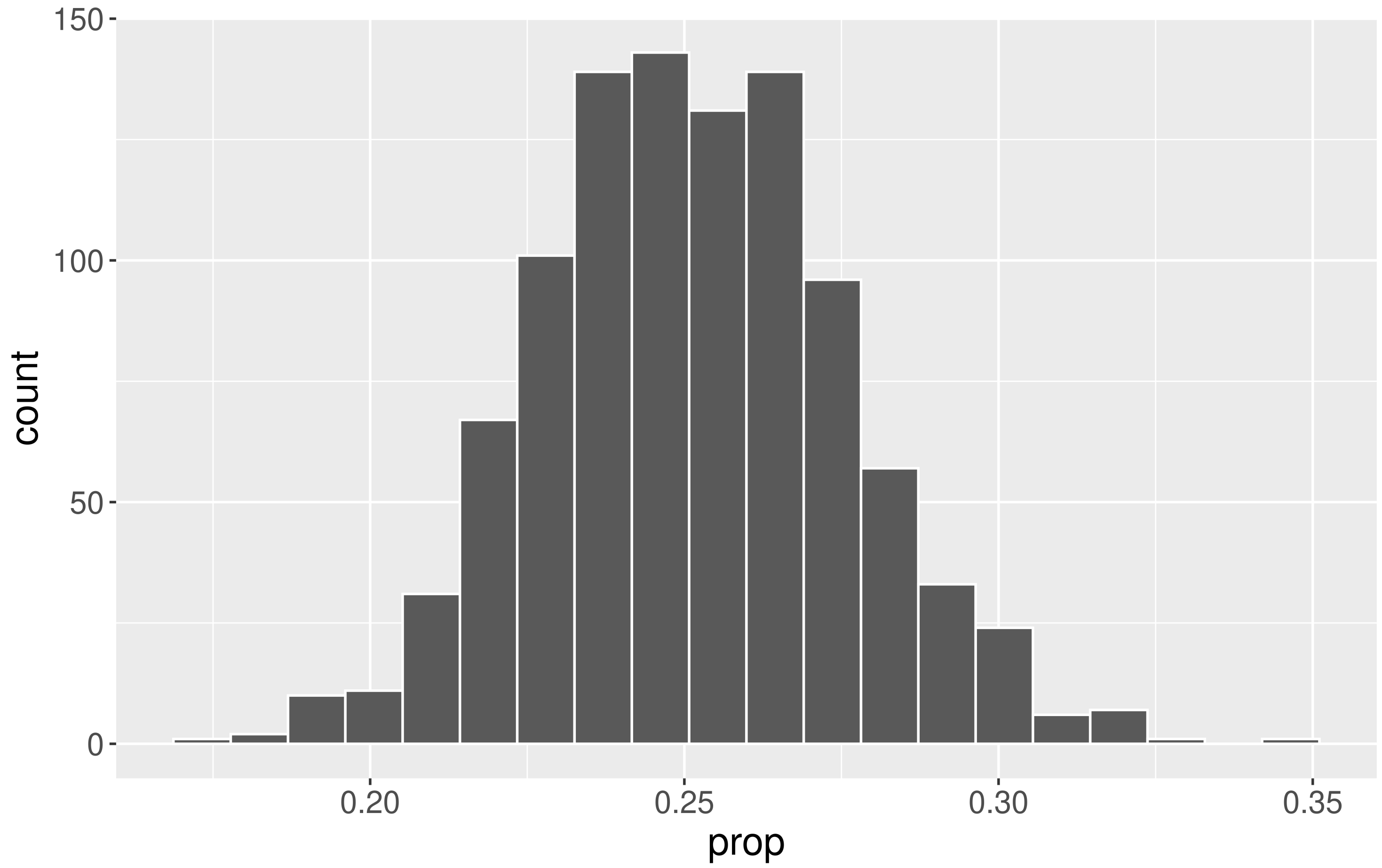
5. Repeat Steps (2) - (4) many (1000+) times.

```r
guess_sampling_dist <- do(1000)*rflip(n = 329, prob = 0.25)
guess_sampling_dist
```

```
       n heads tails      prop
1    329    82   247 0.2492401
2    329    70   259 0.2127660
3    329    71   258 0.2158055
4    329    74   255 0.2249240
5    329    96   233 0.2917933
6    329    83   246 0.2522796
7    329    92   237 0.2796353
8    329    87   242 0.2644377
9    329    83   246 0.2522796
10   329    86   243 0.2613982
11   329    88   241 0.2674772
12   329    82   247 0.2492401
13   329    87   242 0.2644377
14   329    79   250 0.2401216
```

Sampling Distribution Under No ESP

```r
ggplot(data = guess_sampling_dist,
       mapping = aes(x = prop)) +
  geom_histogram(color = "white",
                 bins = 20)
```

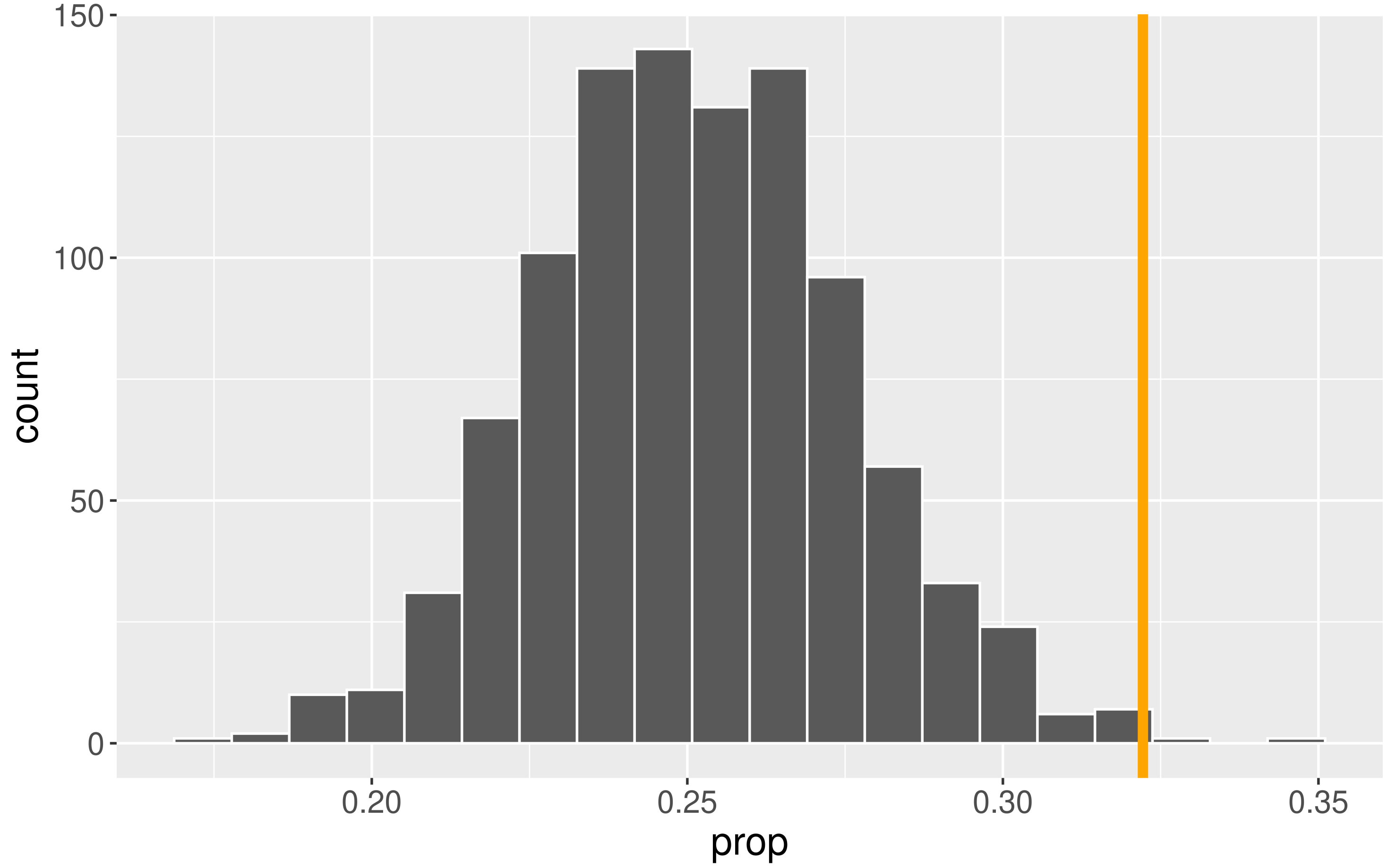


What value should our sampling distribution be centered around if the receivers are just guessing?

Sampling Distribution Under No ESP

- How do the study results compare to the sampling distribution under no ESP?
  - How unusual is it to guess correctly 106 out of 329 times if ESP doesn’t exist?

```r
p_hat <- 106/329
ggplot(data = guess_sampling_dist,
       mapping = aes(x = prop)) +
  geom_histogram(color = "white",
                 bins = 20) +
  geom_vline(xintercept = p_hat,
             color = "orange",
             size = 2)
```

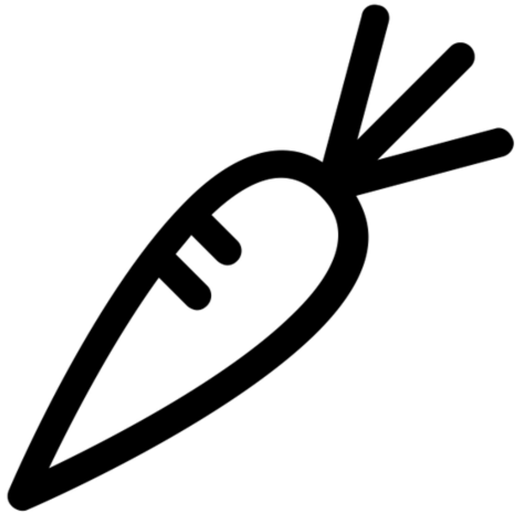


- Do Bem and Honorton have evidence that ESP exists?

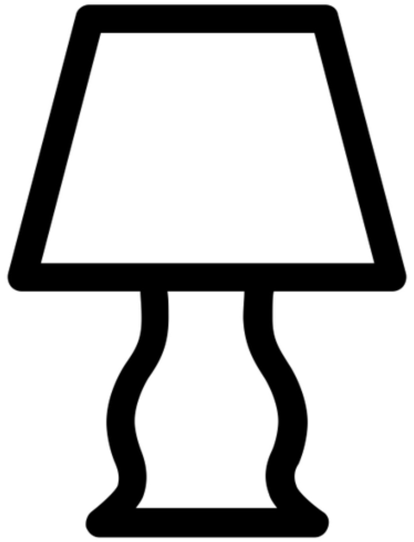
Do Harvardians Have ESP?

In pairs:

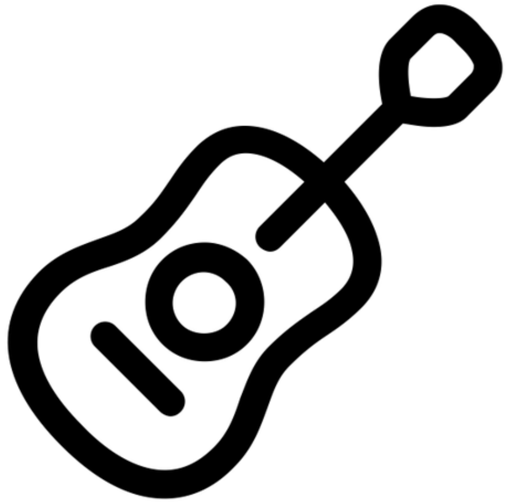
- Decide who is going to be the sender and who is going to be the receiver.
- Sender: Think of one of these images.
- Receiver: Guess which image the sender was thinking of.
- Now switch roles and do it again!
- Once you have both played each role, each person should add a tally mark on the chalkboard.



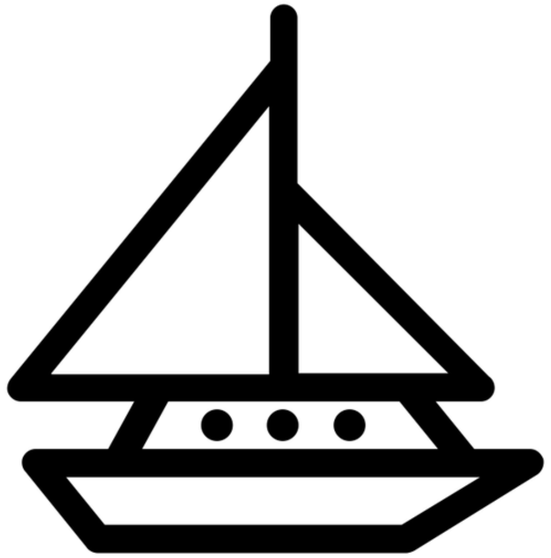
Created by Focus Lab
from Noun Project



Created by Focus Lab
from Noun Project



Created by Focus Lab
from Noun Project



Created by Focus Lab
from Noun Project

Do Harvardians Have ESP?

What do we need to modify in the code to answer the question?

```
guess_sampling_dist <- do(1000)*rflip(n = 329, prob = 0.25)
p_hat <- 106/329
ggplot(data = guess_sampling_dist, mapping = aes(x = prop)) +
  geom_histogram(color = "white", bins = 20) +
  geom_vline(xintercept = p_hat, color = "orange", size = 2)
```

Hypothesis Testing

Big Idea:

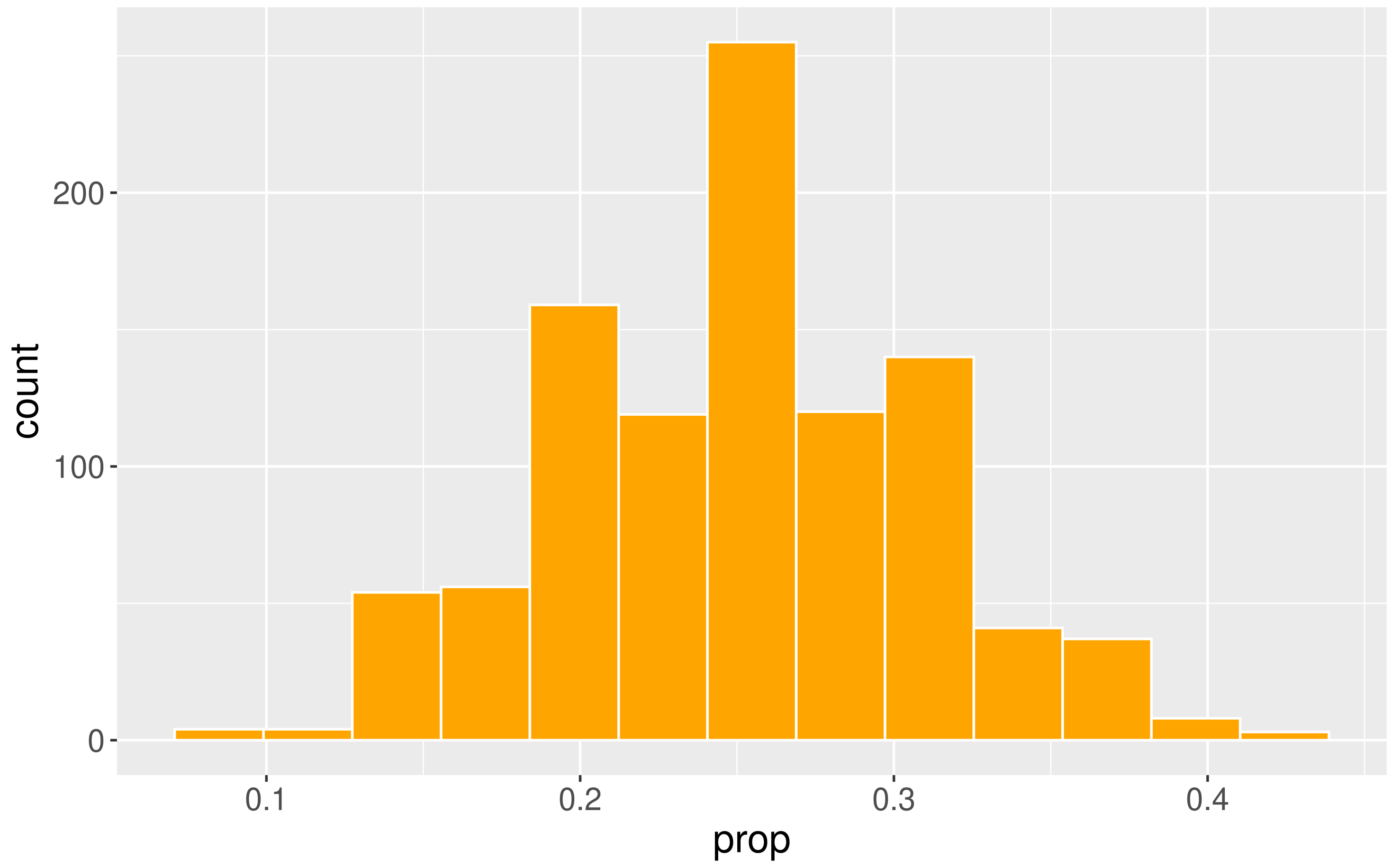
- Make an assumption about the population **parameter**.
- Generate a sampling distribution for a *test* statistic based on that assumption.
  - Called a **null distribution**
- See if the test statistic based on the observed sample aligns with the generated sampling distribution or not.
- If it does, then we didn't learn much.
  - (Didn't prove the parameter equals the assumed value but it is still plausible)
- If it doesn't, then we have evidence that our assumption about the parameter was wrong.

ESP Example

Big Idea:

- Make an assumption about the population parameter.
    - Ex: ESP doesn't exist. p, probability of guessing correctly, equals 0.25.

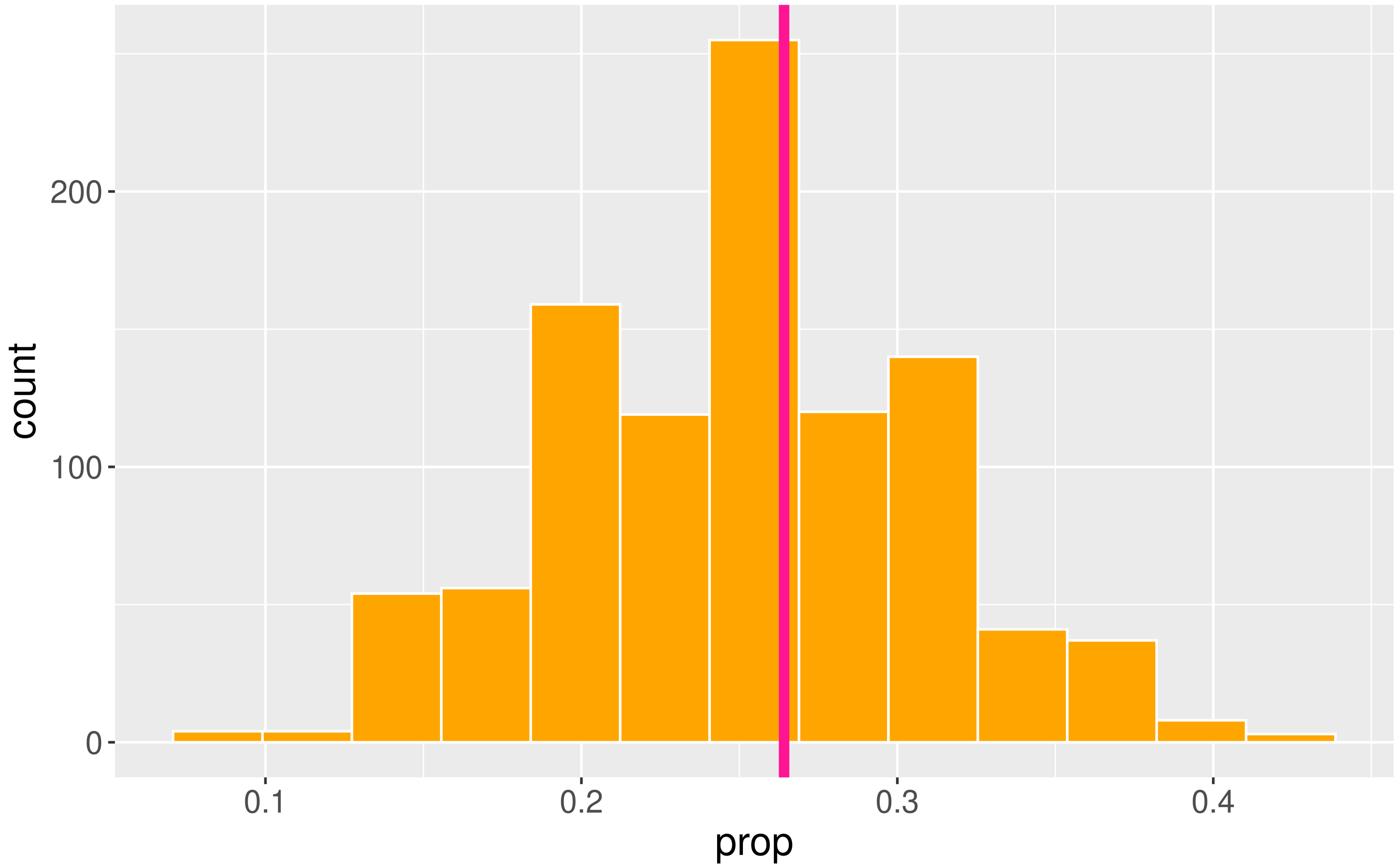
- Generate a sampling distribution for a *test* statistic based on that assumption.
    - Called a **null distribution**



ESP Example

Big Idea:

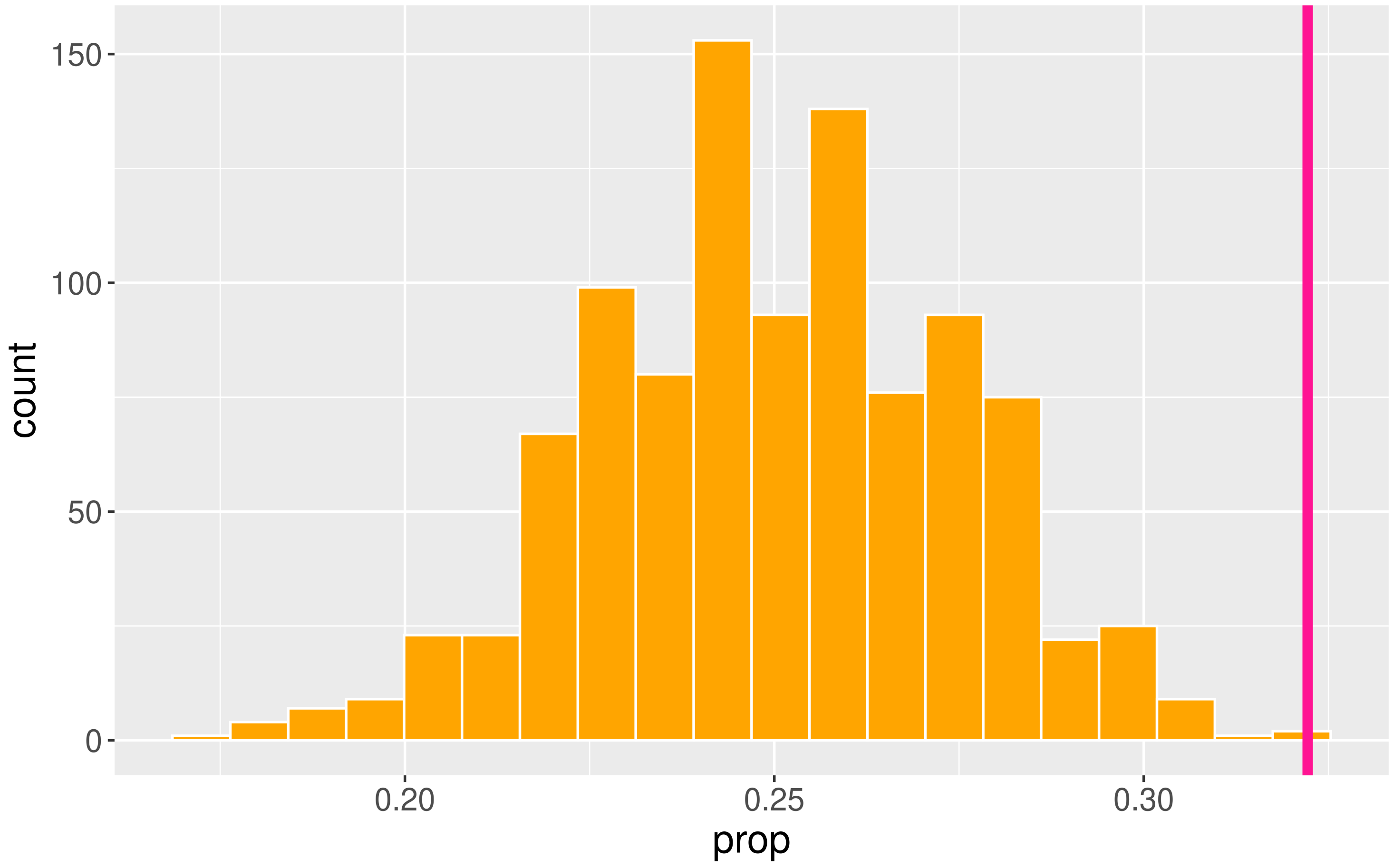
- See if the test statistic based on the observed sample aligns with the generated sampling distribution or not.
  - Ex: It is in the center-ish of the distribution. It isn't an unusual value.
- If it does, then we didn't learn much. (Didn't prove the parameter equals the assumed value but it is still plausible)
  - It is still possible that ESP doesn't exist.



ESP Example

Big Idea:

- See if the test statistic based on the observed sample aligns with the generated sampling distribution or not.
  - It is far in the tails of the distribution. It is an unusual value.
- If it doesn't, then we have evidence that our assumption about the parameter was wrong.
  - We have evidence ESP exists.



Let’s Take a Step Back from Our Last Statement…

- Two important words in data analysis:
  - Reproducibility
  - Replicability

- Reproducibility: If I give you the raw data and my write-up, you will get to the exact same final numbers that I did.
- By using `Quarto` Documents, we are learning a reproducible workflow.
- Replicability: If you follow my study design but collect new data (i.e. repeat my study on new subjects), you will come to the same conclusions that I did.

Replication Crisis

- Science is going through a replication crisis right now.
  - In cancer science, many "discoveries" don't hold up
  - Estimating the reproducibility of psychological science
  - Psychology Is Starting To Deal With Its Replication Problem
- And, sadly, replication studies of Bem and Honorton's ESP trials typically failed to find evidence of ESP.

Reminders:

- Don’t forget that the midterm exam rewrites are due on Thursday at 5pm on Gradescope.
  - Make sure to use the Quarto doc in the Midterm Exam (Rewrites) project on Posit Cloud.
- 🎉 We are now accepting Course Assistant/Teaching Fellow applications for Stat 100 for next semester. To apply, fill out this application by **Nov 15th**.
  - About 10-12 hours of work per week.
  - Primary responsibilities: Attend weekly team meetings, lead a discussion section, hold office hours, grade assessments.